Vol.2015-NL-220 No.9
2015/1/20

観点情報を用いた行列分解によるマルチラベル文書の分類

丸田 要[1]　永井 秀利[1]　中村 貞吾[1]

概要：テキスト分類には分類を行うユーザの目的・観点により分類結果が異なるという性質が存在している．つまり，ある単一の文書データは観点が異なると分類されるクラスが異なる場合がある．その場合ユーザが考える分類とシステムによる分類に差異ができ，その差異部分に含まれる文書データはユーザの情報検索の阻害や見落としを発生させると考えられる．そこで，ユーザによる文書分類例から観点を抽出し，その観点情報をテキスト分類に反映させることでユーザの望む分類を行う．それにより，ユーザが目的の文書を効率よく検索することができることを目指す．本論文では，テキスト分類手法として NMF を含む行列分解を利用するが，その際にテキスト分類に反映させる観点情報の適用方法を複数提案する．そして，実験による比較により各適用方法を評価する．

1. はじめに

インターネットが普及して以来，多くの人が気軽にインターネットを利用して情報収集することができるようになった．また，文書群である記事もネット上に膨大な量が存在している．そのため，膨大な量の文書集合から効率よく目的の文書を検索する手法が必要である．

そして現在，文書集合を効率良く整理・検索する手法の一つとしてクラスタリング検索手法がある．この手法は検索結果をクラスタリングし分類することで目的のカテゴリに絞って目的の文書を探すことができる．しかし，テキスト分類には分類を行うユーザの目的・観点により分類結果が異なるという性質が存在している．つまり，ある単一の文書データは観点が異なると分類されるクラスが異なる場合がある．その様に複数のクラスに属する可能性のあるデータをマルチラベル文書と呼ぶ．マルチラベル文書データ同士は類似度が高い場合が多く分類が困難である．その場合ユーザが考える分類とシステムによる分類に差異ができ，その差異部分に含まれる文書データはユーザの情報検索の阻害や見落としを発生させると考えられる．そこで，ユーザによる文書分類例から観点を抽出し，その観点情報をテキスト分類に反映させることでユーザの望む分類を行う．それにより，ユーザが目的の文書を効率よく検索することができることを目指す．

本論文では，観点はユーザが分類した教師文書中の単語の出現頻度に現れると考える．そこで，観点情報をユーザの観点により分類された教師文書の文書ベクトルから抽出する．その文書ベクトルは単語の出現頻度を特徴量とする．

実際に教師文書ベクトルから観点情報を抽出する手法を 4 つ提案する．各抽出手法はクラスごとに観点情報の抽出を行い，全てのクラスの観点情報を合わせて利用することで観点を近似的に表現する．一つ目の抽出法は，文書ベクトルの平均特徴量を観点情報とする．二つ目の抽出法は，所属クラスの平均特徴量と非所属クラスの平均特徴量の比率を観点情報とする．三つ目の抽出法は，所属クラスの最大特徴量を観点情報とする．四つ目の抽出法は，所属クラスの最大特徴量と非所属クラスの最大特徴量の比率を観点情報とする．

観点情報の抽出が出来たらその観点情報を文書分類に反映させるが，我々は文書分類に行列分解を利用する．そのため，抽出した観点情報を行列で表現し反映させる．

文書分類に用いる行列分解手法は単純な行列分解を利用した手法と NMF[1][2] を利用した手法の二種類の手法を用いて比較を行う．両方の手法において各文書と各クラスとの関連度を表している特徴行列を求める．そして，関連度が最大であるクラスに各文書を分類する．

単純な行列分解を利用する手法では，文書行列を基底行列と特徴行列に分解する行列分解式を利用する．その分解式から基底行列と文書行列から特徴行列を求める．

NMF を利用した手法では，既存の NMF により特徴行列を求める．また，教師あり NMF である NMF-I[3] による手法との比較も行う．

上記の二種類の行列分解手法に抽出して作成した観点行列を導入することでユーザの観点に沿った分類を目指す．

実験では，実際に文書データを分類することで，観点情

[1] 九州工業大学

報の各抽出法と観点行列を導入した各行列分解手法による文書分類の結果を比較する．

2. 観点情報の利用

ユーザの観点は人の感覚に依るところが大きいため，ユーザが文書分類を行う際の観点を明示的に表現することは困難である．そこで，ユーザが分類した教師文書から観点の特徴を抽出し，それを分類に反映することを考える．本論文では観点の特徴は教師文書中の単語の分布に現れると考え，各クラスの単語ごとに観点に対する寄与度を定めることにより近似的に観点を表現する．

本論文では，観点はユーザが分類した教師文書中の単語特徴に現れると考える．そこで，観点情報をユーザの観点により分類された教師文書の文書ベクトルから抽出する．その文書ベクトルは単語の出現頻度を特徴量とする．各クラスの各単語ごとに観点に対する寄与度を定め，全ての寄与度を合わせて利用することにより近似的に観点を表現する．以降では観点に対する寄与度を観点度と呼ぶ．

我々は教師文書ベクトルから観点情報を抽出する手法を4つ提案する．各抽出手法は各クラスにおける単語ごとの観点度を求める手法である．各抽出法の説明では任意のクラス A における単語 t の観点度の抽出法について述べる．クラス A 所属の文書ベクトル集合を D_A とし，それ以外のクラス所属の文書ベクトル集合を $D_{\overline{A}}$ とする．全単語の観点度を求めベクトル化したものを観点単語ベクトルとする．

2.1 観点抽出法 1 ー 平均

観点抽出法 1 は，単語 t に対する D_A の各特徴量において平均をクラス A における単語 t の観点度として求める．この手法は各クラスの平均的な特徴をとることができるため，クラス内のあらゆる文書データに含まれる単語の観点度が高くなることが期待される．

2.2 観点抽出法 2 ー 平均の比率

観点抽出法 2 は，単語 t に対する D_A の各特徴量における平均特徴量 f_A と単語 t に対する $D_{\overline{A}}$ の各特徴量における平均特徴量 $f_{\overline{A}}$ との比率 $f_A/f_{\overline{A}}$ をクラス A における単語 t の観点度として求める．この手法は他のクラス文書 $D_{\overline{A}}$ でも高い特徴量の単語は観点度が低くなり，他のクラス文書 $D_{\overline{A}}$ では低い特徴量の単語は観点度が高くなることが期待される．

2.3 観点抽出法 3 ー 最大値

観点抽出法 3 は，単語 t に対する D_A の各特徴量において最大値をクラス A における単語 t の観点度として求める．この手法はクラス A 文書 D_A 内における高々一文書内でのみ使用される単語でもその文書内で特徴量が高ければ，その単語は観点度が高くなると考えている．

2.4 観点抽出法 4 ー 最大値の比率

観点抽出法 4 は，単語 t に対する D_A の各特徴量における最大値特徴量 max_A と単語 t に対する $D_{\overline{A}}$ の各特徴量における最大値特徴量 $max_{\overline{A}}$ との比率 $max_A/max_{\overline{A}}$ をクラス A における単語 t の観点度として求める．この手法は観点抽出法 2 と 3 を合わせた効果が期待される．

3. 関連研究

3.1 NMF

本論文では，最近文書分類で注目を浴びている非負値行列因子分解 (NMF)[2] を行列分解による文書分類の一手法として利用しているため，この 3 節では NMF について述べる．NMF は，文書クラスタリング手法の一つであり，高次元でスパースな文書行列をクラスタリングするのに適している．

NMF は式 (1) のように m 個の文書データと n 個の索引語から作られる $n \times m$ の索引語文書行列 X を $n \times k$ の基底行列 U と $k \times m$ の特徴行列 V^T の積の形に分解することにより文書データを次元圧縮することができる．その次元圧縮結果である特徴行列 V が各文書と各クラスタとの関連度を表している．ここで，k はクラスタ数である．

$$X = UV^T \tag{1}$$

つまり，n 次元の文書データである索引語文書行列 X が k 次元の文書データである特徴行列 V^T へと次元圧縮される．NMF を文書クラスタリングへ適用する際には次元圧縮後行列である特徴行列 V^T を利用する．特徴行列 V^T の h 行目の要素の値が，各文書と h 番目のクラスタとの関連度の大きさを表している．そのため，i 番目の文書データのクラスタは (2) 式で得られる．

$$\text{文書 i のクラス} = \arg\max_h \ v_{ih} \tag{2}$$

基底行列 U と特徴行列 V への分解は NMF の目的関数である式 (3) の J を最小にするような基底行列 U と特徴行列 V を推定することで求まる．

$$J = ||X - UV^T||^2 \tag{3}$$

そして，ラグランジュの未定乗数法を用いて式 (3) の J を最小にする基底行列 U と特徴行列 V の乗算型更新式を求める．r を反復更新の更新回数として式 (4)，(5) のように表される．

$$v_{ij}^{(r+1)} \leftarrow v_{ij}^{(r)} \frac{(X^TU)_{ij}}{(VU^TU)_{ij}} \tag{4}$$

$$u_{ij}^{(r+1)} \leftarrow u_{ij}^{(r)} \frac{(XV)_{ij}}{(UV^TV)_{ij}} \tag{5}$$

ここで，$u_{ij}^{(r)}$ と $v_{ij}^{(r)}$ はそれぞれ更新回数 r 回目である U と V の i 行 j 列の要素を表し，$(X)_{ij}$ は行列 X の i 行 j 列

の要素を表す．

また，各繰り返し後には発散を防ぐためと各基底を単位ベクトルにするために基底行列 U を以下の式 (6) に従い正規化を行う．

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \tag{6}$$

通常，基底行列 U と特徴行列 V の初期値 $U^{(0)}$ と $V^{(0)}$ はランダムな値を与えることで作成される．

3.1.1 初期値に関する問題点

NMF では，$U^{(0)}$ と $V^{(0)}$ の初期値によって，最終的に得られる $U^{(R)}$ と $V^{(R)}$ は異なる．ここで R は最大更新回数とする．つまり，$V^{(R)}$ はクラスタリング結果を表しているため，クラスタリング結果は初期値 $U^{(0)}$ と $V^{(0)}$ に依存していると言える．

3.2 NMF-I

NMF には 3.1.1 節で挙げたように収束結果が初期値 $U^{(0)}$ と $V^{(0)}$ に依存するという問題が存在する．一般的には NMF での初期値は乱数で与えるが，単純な乱数ではクラスタリング結果が悪い局所解に収束するような初期値となる可能性がある．

そこで，我々はその問題に対応した NMF-I[3] を前の論文で提案している．NMF-I では，既知である教師文書ベクトルの各クラスタにおける平均ベクトルを求め，それを平均教師ベクトルとする．その平均教師ベクトルを各クラスタ毎に並べ基底とした教師基底行列 U_s を考えた場合，NMF での理想的な文書分類がなされた場合の基底行列の収束値は U_s に近いものであろうとの期待に基づき，この教師基底行列 U_s を教師あり NMF における基底行列 U の初期値とする．教師基底行列 U_s は式 (7) で与える．

$$U_s = X_{train}(V_{train}^T)^+ \tag{7}$$

ここで，教師データ数を t とした時，$X_{train}(n$ 行 t 列) は教師データのみの索引語行列であり，$V_{train}^T(k$ 行 t 列) は各文書の正解クラスタに対応する要素を 1 としてそれ以外の要素を 0 とした行列である．また，“+” は疑似逆行列である．

教師基底行列 U_s を U の初期値 $U^{(0)}$ として，以降は既存 NMF と同様の更新式を実行する手法を NMF-I(NMF with Initial basis value by traing data) と呼ぶことにしている．

4. 行列分解による文書分類

観点情報の抽出が出来たら，その観点情報を文書分類に反映させる．その際の文書分類には行列分解を利用する．そのため，抽出した観点情報を行列空間表現として整理する．整理して出来た行列を観点行列 U_m として文書分類に反映させる．観点行列 U_m は 2 節の各抽出法における各クラスごとの観点単語ベクトルを全てのクラス分並べて作成する．

文書分類に用いる行列分解手法は式 (1) を特徴行列 V^T について求めるために整理した式を利用する手法と NMF[1][2] を利用する手法の二種類の手法を用いて比較を行う．両方の手法において各文書と各クラスとの関連度を表している特徴行列 V^T を求める．そして，関連度が最大であるクラスに各文書を分類する．

式 (1) の行列分解式を利用する手法では，行列分解式から基底行列 U と文書行列 X を与えて特徴行列 V^T を求める．

NMF を利用した手法では，教師あり NMF の一手法である NMF-I[3] により特徴行列 V^T を求める．

上記の二種類の行列分解手法に抽出して作成した観点行列 U_m を導入することでユーザの観点に沿った分類を目指す．

4.1 分類手法 A ー 行列分解

分類手法 A では，NMF でも使用する式 (1) を利用する．式 (1) では索引語文書行列 X を基底行列 U と特徴行列 V^T の積の形に行列分解している．実際に行列分解する際に基底行列 U と特徴行列 V^T が与えられていない場合は NMF により行列分解を行う．しかし，2 節で求める観点行列を利用することで NMF を用いずとも行列分解が可能となる．

具体的に分類手法 A は，式 (1) における基底行列 U に観点行列 U_m を代入し特徴行列 V^T を解く式 (8) の形に整理することで特徴行列 V^T を求める．

$$V^T = U_m^+ X \tag{8}$$

ここで，式 (8) において A^+ は行列 A の疑似逆行列である．

4.2 分類手法 B ー 行列分解

分類手法 A での分類結果は観点行列 U_m の値でほとんど決定してしまう．さらに，教師文書データから観点行列 U_m は作成しているため，教師文書データに特異なデータがあるときテスト文書データの最適な観点情報や分類結果が得られるとは言えないという問題がある．

そこで，観点行列を最適な結果へと修正するために式 (8) に NMF による文書分類で求めた基底行列 U を導入する．それにより，教師データのみから作成した観点行列よりも最適な基底行列へと修正された行列分解による文書分解が行えると期待する．

つまり，分類手法 B は NMF で求めた基底行列 U を式 (8) の観点行列 U_m に追加した式 (9) で特徴行列 V^T を求める．

$$V^T = (\mu U_m + U)^+ X \tag{9}$$

ここで，式 (9) において μ は観点行列の重み，A^+ は行

列 A の疑似逆行列である．

NMF で求める基底行列 U は各クラスにおける各単語に対する重要度とみなす事ができる．そこで，この基底行列 U に観点行列 U_m を加算することは観点を表すであろう単語の重要度を強調することであると考える．

4.3 分類手法 C ー NMF-I

分類手法 C では 3.2 節で述べた NMF-I を利用する．実際には，NMF-I の教師基底行列 U_s の代わりに 2 節で求めた観点行列 U_m を代入する．それにより，観点行列 U_m を初期値とした更新式による柔軟な収束ができ分類手法 A や分類手法 B よりも最適な分類結果に収束することが期待される．

5. 実験

実際に教師文書集合から観点抽出法 1〜4 までの方法で観点行列 U_m を作成する．そして，既存 NMF，NMF-I，分類手法 A，分類手法 B，分類手法 C で分類し比較することで，提案手法の有効性を検証する．

5.1 実験用文書データセット

本論文ではシングルラベル文書とマルチラベル文書の文書データセットを用いて比較を行っている．シングルラベルの文書データは正解となるクラスが 1 つである文書データである．マルチラベルの文書データは正解となるクラスが複数である文書データである．シングルラベル文書データセットとして CLUTO のサイト[*1] で公開されている文書集合を，マルチラベル文書データセットとしてロイターニュース[*2] で公開されている文書集合を利用する．各シングルラベルの文書データセットの詳細は表 1 に示す．データセットの k1a, k1b と wap は Yahoo!内の様々な web ページから構成され，re0 はロイターのニュースワイヤーから取得したニュース記事で構成されている．また，tr31，tr41 は TREC のテスト用文書である．そして，fbis は米国の元政府機関である Foregin Broadcast Information Service (FBIS) が収集したニュース記事で構成されている．各マルチラベルの文書データセットの詳細は表 2 に示す．本論文においては観点の違いを簡単に比較するために 2 クラスのマルチラベル文書を対象に実験を行う．表 1，2 においてクラス間類似度とは各クラスの重心のコサイン類似度である．

表 2 のマルチラベル文書データの分類実験では複数の観点により分類を行い観点の違いを比較する．文書データ reut2-gc は 3 つの観点で，reut2-cs と reut2-tb は 2 つの観点により分類を行う．観点の違いを明確にするために正解ラベル (クラス) を変化させている．そして，本来なら人手によるラベリングによる観点の違いを設定することが望ましいが，本論文の観点の違いは機械的にランダムで設定している．表 3 に観点の違いによるラベルの変化の割合を示す．

[*1] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download/
[*2] http://www.daviddlewis.com/resources/testcollections/reuters21578/

表 1 シングルラベルの文書データセット

データ名	文書数	索引語数	クラス数	クラス間類似度
k1a	2340	21839	20	0.219
k1b	2340	21839	6	0.220
re0	1504	2886	13	0.276
wap	1560	6460	20	0.212
tr31	927	10128	7	0.191
tr41	878	7454	10	0.171
fbis	2463	2000	17	0.252

表 2 マルチラベルの文書データセット

データ名	文書数	索引語数	クラス数	クラス間類似度
regc(観点 1)	242	3364	2	0.808
regc(観点 2)	242	3364	2	0.824
regc(観点 3)	242	3364	2	0.790
recs(観点 1)	74	1908	2	0.601
recs(観点 2)	74	1908	2	0.601
retb(観点 1)	60	1164	2	0.658
retb(観点 1)	60	1164	2	0.666

表 3 観点の違いによる正解ラベルの変化

データ名	比較観点	文書数	変化数	変化割合
regc	観点 1 vs 観点 2	242	52	21.5%
regc	観点 1 vs 観点 3	242	48	19.8%
regc	観点 2 vs 観点 3	242	51	21.1%
recs	観点 1 vs 観点 2	74	17	23.0%
retb	観点 1 vs 観点 2	60	12	20.0%

5.2 評価方法

分類結果の評価値には Entropy，Purity，RandIndex，Precision，Recall 及び F 値 (Precision と Recall の調和平均) を用いる．さらに最終的な分類性能値はこれらの F 値を除く五種類の評価値の調和平均 Hm により評価する．

Entropy は式 (10) より求める．Entropy は各クラスタにおける正解集合の分布度合を表しており，小さな値ほどクラスタリング結果が良好であることを意味している．ここで N は総文書数を示す．また，調和平均 Hm を算出する際には $(1 - Entropy)$ として計算する．

$$Entropy = \sum_{i=1}^{k} \frac{|C_i|}{N} \times (-\sum_{h=1}^{k} P(A_h|C_i) \log P(A_h|C_i)) \quad (10)$$

Purity は結果クラスタに一番多く含まれている正解クラスタを用いて，結果クラスタに正解データが含まれている割合を示す指標である．クラスタリング結果の Purity は，

各クラスタのデータ数による重み付き平均をとるように定義し，式 (11) に示す．

$$Purity = \frac{1}{N}\sum_{i=1}^{k}\max_{h}|C_i \cap A_h| \tag{11}$$

式 (10)，式 (11) において C_i はクラスタリング結果に対する i 番目の文書のクラスタであり，A_h は正解データに対する h 番目の文書のクラスタである．$A_h \cap C_i$ は正解データであるクラスタ A_h とクラスタリング結果のクラスタ C_i が共通している文書数である．

RandIndex はデータの各ペア同士の正解が同じクラスタならば同じクラスタになるかどうかの判定の正解率を表し式 (12) で求める．

$$RandIndex = \frac{TP+TN}{TP+FP+FN+TN} \tag{12}$$

ここで TP は同じ正解クラスタであるデータのペアが結果クラスタで同じクラスタである対の数，TN は異なる正解クラスタであるデータのペアが結果クラスタで異なるクラスタである対の数，FP は異なる正解クラスタであるデータのペアが結果クラスタで同じクラスタである対の数，FN は同じ正解クラスタであるデータのペアが結果クラスタで異なるクラスタである対の数を表す．

Precision はクラスタリング結果の中にどの程度正解が含まれているを表す．実際には以下の式 (13) で求める．

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

Recall は正解データがどの程度結果クラスタで正しくクラスタリングされているかを表す．実際には以下の式 (14) で求める．

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

5.3 実験方法及び結果

実験では，20 種類の異なる教師集合と初期値 $U^{(0)}$ と $V^{(0)}$ を準備する．それらの教師集合と初期値に対して各手法で文書分類を行い平均評価値を調査した．実験における μ は 1 とする．

NMF における更新回数は 50 回とする．各文書データセットの文書数に対して各クラスの 15 データを教師データとして使用する．分類結果を評価したものを表 7〜14 に示す．実験結果では分類手法を分類と省略する．また，平均を利用した観点抽出法 1 が NMF-I における U_s の作成法と同等であるため表 7 と表 11 での「NMF-I」と「分類 C」の分類結果は同じである．表 7〜14 の値に付随する「*」は各文書データを分類した際の最も良い評価値を示している．

6. 考察

6.1 シングルラベル文書の分類についての考察

シングルラベル文書の分類結果を比較する．表 7〜10 の結果から各観点抽出法を使用する際の最も相性が良い分類手法が考えられる．

まず，観点抽出法 1 では行列分解を利用した分類手法 A が良い分類性能である．そのため観点抽出法 1 と相性がよい手法は分類手法 A であると考えられる．

次に，観点抽出法 2〜4 ではほとんどの評価値において分類手法 C が最もよい分類性能であることがわかる．そのため観点抽出法 2〜4 と相性がよい手法は NMF-I を利用した分類手法 C であると考えられる．

また，表 7 における分類手法 A による文書データ k1b，wap と tr31 の 3 つの評価値は他の手法を合わせて考えた最大評価値と同程度である．そして，他の文書データの評価値は観点抽出法 1 を利用した分類手法 A が最大である．

以上のことから，実験に使用したシングルラベル文書の分類には観点抽出法 1 と分類手法 A を組み合わせた手法がよい分類性能であると考えられる．

6.2 マルチラベル文書の分類についての考察

表 11〜14 において文書データ regc はどの観点であっても他のマルチラベル文書の分類結果よりも評価値が比較的に低い．これは表 2 に示されているように regc のクラスタ間類似度が他の文書データよりも高いため分類が困難であったためと考えられる．また，シングルラベル文書よりもマルチラベル文書の方が全体的に評価値が低い理由はシングルラベル文書のクラスタ間類似度よりマルチラベル文書のクラスタ間類似度が高いことが原因であると考えられる．

6.3 観点抽出法と分類手法との組合せ (マルチラベル)

マルチラベル文書の分類結果から各観点抽出法と分類手法の組合せを比較し考察する．表 11〜14 の評価値を比較して相性を整理したものを表 4 に示す．

表 4 各分類手法と各観点抽出法との相性順位

	分類手法 A	分類手法 B	分類手法 C
抽出法 1	3 位	4 位	3 位
抽出法 2	3 位	4 位	2 位
抽出法 3	1 位	1 位	4 位
抽出法 4	1 位	2 位	1 位

表 4 における値は各分類手法に着目した際の分類手法との相性の順位を表している．

順位の決め方は着目する分類手法ごとに決める．各抽出法において各文書データの最大 (または最大と同程度) の評価値が存在する個数の多い順で相性順位を決定する．

分類手法 A に着目して各抽出法を比較すると抽出法 3 と 4 において評価値が最大と同等である文書データがそれぞれ 4 つ存在する．そのため分類手法 A と相性がよいのは抽出法 3 と 4 であると考えられる．

次に分類手法 B に着目する．分類手法 B では評価値が最大か最大と同程度である文書データの数が抽出法 3 で 4 つと一番多い．そのため分類手法 B と相性がよいのは抽出法 3 であると考えられる．

そして分類手法 C に着目すると抽出法 4 が他の抽出法より評価値が高い．そのため分類手法 C と相性がよいのは抽出法 4 であると考えられる．

6.4 観点の違いによる影響

6.4.1 観点行列への影響

マルチラベル文書では観点が異なることにより各文書の正解ラベルが異なる．そのため，観点行列を作成する際に利用する教師文書データの正解ラベルも異なる．つまり，観点の違いが観点行列に影響を与える．ある観点 $v1$ における観点行列とそれ以外の観点 $v2$ における観点行列の非類似度を表 5 に示す．観点行列同士の非類似度が低いということは観点の違いによる影響が小さく，高い場合は観点の違いによる影響が大きいと考えられる．

表 5 より観点抽出法 2 が最も観点の違いによる影響が現れていると考えられる．

また，観点抽出法 2 と観点抽出法 4 が他 2 種の方法より非類似度が大きい．このことより，他クラスとの比率を利用することは観点の影響が反映されやすいと考えられる．

表 5 観点行列の非類似度

Data ($v1$ vs $v2$)	抽出 1	抽出 2	抽出 3	抽出 4
regc(観点 1 vs 観点 2)	0.009	**1.874**	0.401	**1.875**
regc(観点 1 vs 観点 3)	0.008	**1.840**	0.453	**1.840**
regc(観点 2 vs 観点 3)	0.009	**1.871**	0.458	1.868
recs(観点 1 vs 観点 2)	0.007	**1.783**	0.411	1.770
retb(観点 1 vs 観点 2)	0.007	**1.701**	0.361	1.670

6.4.2 ラベル変化文書の正解率への影響

前述したとおり観点が異なると正解ラベルに違いが出てくる．そこで，正解ラベルが観点によって異なるデータのみの正解率を表 15〜18 に示す．表 15〜18 の NMF の欄では既存 NMF における各観点の正解率の平均を示している．そして，提案した分類手法の欄では既存 NMF における正解率からの増加割合を示している．つまり，この表の増加率が高いと観点を変更した際に観点の違いを考慮した各抽出法と各分類手法ができていることを表している．

表 15〜18 において各分類手法に着目した際の各抽出法との相性順位を表 6 に示す．順位の決定法は表 4 の際と同等である．

6.5 マルチラベル文書分類について

表 4 と表 11〜14 の結果はラベル変化文書以外の分類結果の影響が含まれている．しかも，ラベル変化率は高々 20 数%しかない．そのため，それ以外の影響が大きいと考えられる．つまり，表 4 と表 11〜14 の結果からは全体の分類性能は推測できるが，観点の違いに対する影響が推測出来ない．

表 6 各分類手法と各観点抽出法との相性順位 - ラベル変化文書

	分類手法 A	分類手法 B	分類手法 C
抽出法 1	3 位	3 位	4 位
抽出法 2	4 位	4 位	2 位
抽出法 3	1 位	1 位	4 位
抽出法 4	1 位	1 位	1 位

そこで，表 4 と表 15〜18 を含めてマルチラベル文書の分類結果を比較する．表 4 と表 15〜18 の結果はラベル変化文書のみの結果であるため観点の違いに対する影響が推測できる．

全体の結果とラベル変化文書のみの結果はどちらも良い結果であることが望ましいので，表 4 と表 6 より分類手法 A は観点抽出法 3 と 4，分類手法 B は観点抽出法 3，分類手法 C は観点抽出法 4 がよい組み合わせであると考えられる．

7. おわりに

本論文では，ユーザの観点に沿った分類を目指し，そのための 4 つの観点抽出法と観点情報を利用する 3 つの分類手法を提案した．観点抽出法で観点行列を作成し提案するそれぞれの分類手法に適用させることでユーザの観点に沿った分類を行う．

まず，シングルラベル文書では平均特徴量を観点情報とした観点抽出法 1 と行列分解による分類手法 A との組み合わせが実験中では最も高い分類性能が得られた．

次に，マルチラベル文書では分類結果の評価値にばらつきが見られ一意に最良な組合せを示すことは困難であった．これは，シングルラベル文書よりもマルチラベル文書はクラス間類似度が高いため観点行列や初期値における少しの差が分類結果に影響するためや，現在の実験に用いている観点がシステム的ランダムに依るものであることが原因であると考えられる．しかし，各観点抽出法と各分類手法との大まかな組合せの良し悪しは示すことが出来た．

今後は新たな文書データを導入して性能を調査する必要がある．また，現在はマルチラベル文書は観点による影響をわかりやすくするために 2 クラスを扱っているが，2 クラス以上のデータによる実験と検証が必要である．そして，今は正解ラベルを割り当てる観点をシステムによるランダムとしているが，人手による観点を利用した実験も必要である．

表 7 シングルラベル文書の観点抽出法 1 による Hm

Data	NMF	NMF-I	分類 A	分類 B	分類 C
k1a	0.565	0.727	**0.771***	0.768	0.727
k1b	0.640	0.792	**0.869**	0.752	0.792
re0	0.446	0.506	**0.640***	0.627	0.506
wap	0.527	0.741	0.780	**0.784***	0.741
tr31	0.662	773	**0.900**	**0.900**	0.773
tr41	0.608	775	**0.889***	0.871	0.775
fbis	0.530	639	**0.750***	0.734	0.639

表 8 シングルラベル文書の観点抽出法 2 による Hm

Data	NMF	NMF-I	分類 A	分類 B	分類 C
k1a	0.565	0.727	0.543	0.572	**0.758**
k1b	0.640	0.792	0.809	0.808	**0.870***
re0	0.446	**0.506**	0.418	0.380	0.487
wap	0.527	0.741	0.524	0.577	**0.763**
tr31	0.662	0.773	0.872	0.901	**0.932***
tr41	0.608	0.775	0.751	0.751	**0.847**
fbis	0.530	**0.639**	0.253	0.348	0.620

表 9 シングルラベル文書の観点抽出法 3 による Hm

Data	NMF	NMF-I	分類 A	分類 B	分類 C
k1a	0.565	0.727	0.732	0.731	**0.752**
k1b	0.640	0.792	0.780	0.776	**0.834**
re0	0.446	0.506	0.522	0.521	**0.527**
wap	0.527	0.741	0.746	0.745	**0.767**
tr31	0.662	0.773	0.874	0.872	**0.883**
tr41	0.608	0.775	0.803	0.803	**0.830**
fbis	0.530	0.639	**0.689**	**0.689**	0.661

表 10 シングルラベル文書の観点抽出法 4 による Hm

Data	NMF	NMF-I	分類 A	分類 B	分類 C
k1a	0.565	0.727	0.550	0.547	**0.758**
k1b	0.640	0.792	0.780	0.779	**0.863**
re0	0.446	**0.506**	0.417	0.368	0.476
wap	0.527	0.741	0.554	0.550	**0.762**
tr31	0.662	0.773	0.860	0.860	**0.917**
tr41	0.608	0.775	0.722	0.720	**0.844**
fbis	0.530	**0.639**	0.326	0.335	0.613

表 11 マルチラベル文書の観点抽出法 1 による Hm

Data	NMF	NMF-I	分類 A	分類 B	分類 C
regc(観点 1)	0.027	0.069	**0.083***	0.075	0.069
regc(観点 2)	0.017	0.036	0.061	**0.066**	0.036
regc(観点 3)	0.026	0.086	0.089	**0.108***	0.086
recs(観点 1)	0.091	0.266	**0.383**	0.379	0.266
recs(観点 2)	0.075	0.254	0.373	**0.385***	0.254
retb(観点 1)	0.063	0.276	0.418	**0.440**	0.276
retb(観点 2)	0.079	0.272	0.384	**0.396**	0.272

表 12 マルチラベル文書の観点抽出法 2 による Hm

Data	NMF	NMF-I	分類 A	分類 B	分類 C
regc(観点 1)	0.027	0.069	0.076	**0.077**	**0.077**
regc(観点 2)	0.017	0.036	0.051	0.056	**0.060**
regc(観点 3)	0.026	0.086	0.073	0.080	**0.087**
recs(観点 1)	0.091	0.266	0.375	0.375	**0.388**
recs(観点 2)	0.075	0.254	0.378	**0.379**	0.373
retb(観点 1)	0.063	0.276	0.451	0.463	**0.479***
retb(観点 2)	0.079	0.272	0.392	0.401	**0.404**

表 13 マルチラベル文書の観点抽出法 3 による Hm

Data	NMF	NMF-I	分類 A	分類 B	分類 C
regc(観点 1)	0.027	0.069	**0.077**	**0.077**	0.074
regc(観点 2)	0.017	0.036	0.073	**0.076***	0.063
regc(観点 3)	0.026	0.086	0.093	**0.107***	0.103
recs(観点 1)	0.091	0.266	**0.393***	**0.393***	0.377
recs(観点 2)	0.075	0.254	0.356	0.368	**0.369**
retb(観点 1)	0.063	0.276	**0.459**	0.288	0.415
retb(観点 2)	0.079	0.272	0.276	**0.403**	0.377

表 14 マルチラベル文書の観点抽出法 4 による Hm

Data	NMF	NMF-I	分類 A	分類 B	分類 C
regc(観点 1)	0.027	0.069	**0.084***	0.083	**0.084***
regc(観点 2)	0.017	0.036	0.062	0.062	**0.066**
regc(観点 3)	0.026	0.086	0.080	**0.089**	**0.088**
recs(観点 1)	0.091	0.266	0.375	0.369	**0.383**
recs(観点 2)	0.075	0.254	0.373	0.375	**0.383**
retb(観点 1)	0.063	0.276	0.454	0.465	**0.470**
retb(観点 2)	0.079	0.272	0.408	0.400	**0.411***

表 15 観点によりラベルが異なる文書の正解率 - 観点抽出法 1

Data	比較観点	NMF	分類 A	分類 B	分類 C
regc	観点 1 観点 2	50.0%	**+6.6%***	+4.3%	+3.8%
regc	観点 1 観点 3	52.3%	+3.9%	**+4.8%***	+1.6%
regc	観点 2 観点 3	50.6%	+6.0%	**+6.75%**	+2.9%
recs	観点 1 観点 2	50.0%	**+21.55%***	+21.7%	+12.9%
retb	観点 1 観点 2	50.0%	**+22.4%**	**+22.4%**	+12.8%

表 16 観点によりラベルが異なる文書の正解率 - 観点抽出法 2

Data	比較観点	NMF	分類 A	分類 B	分類 C
regc	観点 1 観点 2	50.0%	**+5.8%**	**+5.8%**	+3.25%
regc	観点 1 観点 3	52.3%	+3.8%	+3.9%	**+4.2%**
regc	観点 2 観点 3	50.6%	+5.8%	**+6.0%**	+5.85%
recs	観点 1 観点 2	50.0%	**+21.35%**	+21.2%	+21.25%
retb	観点 1 観点 2	50.0%	**+25.3%**	+19.2%	+25.2%

表 17 観点によりラベルが異なる文書の正解率 - 観点抽出法 3

Data	比較観点	NMF	分類 A	分類 B	分類 C
regc	観点 1 観点 2	50.0%	**+6.35%**	+6.2%	+4.5%
regc	観点 1 観点 3	52.3%	**+4.35%**	**+4.35%**	+4.25%
regc	観点 2 観点 3	50.6%	**+7.35%***	**+7.35%***	+4.65%
recs	観点 1 観点 2	50.0%	+21.35%	**+21.5%**	+16.85%
retb	観点 1 観点 2	50.0%	**+25.35%**	**+25.35%**	+19.05%

表 18 観点によりラベルが異なる文書の正解率 - 観点抽出法 4

Data	比較観点	NMF	分類 A	分類 B	分類 C
regc	観点 1 観点 2	50.0%	**+6.35%**	+6.2%	+6.25%
regc	観点 1 観点 3	52.3%	**+4.65%**	+4.6%	+4.35%
regc	観点 2 観点 3	50.6%	**+6.15%**	+6.1%	+5.7%
recs	観点 1 観点 2	50.0%	**+21.25%**	+21.0%	+21.1%
retb	観点 1 観点 2	50.0%	**+25.9%***	+25.45%	+25.45%

参考文献

[1] D.D.Lee , H.S.Seung :“Algorithms for Non-negative Matrix Factorization”, NIPS , pp.556-562 , (2000).
[2] W.Xu, X.Liu, and Y.Gong, “Document clustering based on non-negative matrix factorization”, in Proc.ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR), Toronto, ON, Candada, 2003.
[3] 丸田要，永井秀利，中村貞吾，“文書分類のための教師制約を用いた非負値行列因子分解”，情報アクセスシンポジウム 2013，pp14-21，2013.
[4] H.Lee , J.Yoo , S.Choi “Semi-Supervised Nonnegative Matrix Factorization”, IEEE SIGNAL PROCESSING LETTERS , Vol.17 No.1 , pp.4-7 , JANUARY 2010.
[5] 新納浩幸，佐々木稔，“NMF とリンクベースの修正法によるピンポン型文書クラスタリング”，情報処理学会，自然言語処理研究会報告，Vol.2007,no.47,p.7-12.
[6] 新納浩幸，佐々木稔，“Mcut + NMF による文書クラスタリング”，言語処理学会年次大会発表論文集，Vol.13，pp,558-561,(2007).
[7] C.D.Manning, P.Raghavan, H.Schutze, 岩野和生訳，黒川利明訳，濱田誠司訳，村上明子訳，“情報検索の基礎”，共立出版株式会社，2012.
[8] C.Ding,T.Li,W.Peng : “On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing”, Computational Statistics and Data Analysis 52 , 3913 - 3927 (2008).
[9] 堀田政二，宮原末治，“Non-neagtive Matrix Factorization の初期値の設定法とその応用”，電子情報通信学会技術研究報告，Vol.102，No.652，pp.19-24，2003.
[10] 萩野広樹，吉田哲也，“トピックグラフに基づく NMF を用いた転移学習”，IPSJ SIG Technical Report, Vol.2011-MPS-82 No17.
[11] D.Zhang, Z.Zhou, S.Chen, “Nonnegative Matrix Factorization on Kernels”, PRICAI 2006: Trends in Artificial Intelligence Lecture Notes in Computer Science Vol.4099，pp.404-412，2006.
[12] Y.Chen, M.Rege, M.Dong, J.Hua, “Non-negative matrix factorization for semi-supervised data clustering”, Knowl, Inform. Syst., vol.17, pp.355-379, 2008.
[13] F.Wang, T.Li, C.Zhang, “Semi-supervised clustering via matrix factorization”, in Proc. SIAM Int. Conf. Data Mining (SDM), Atlanta, GA, 2008.
[14] C.Ding, T.Li, W.Peng, H.Park, “Orthogonal nonnegative matrix tri-factorizations for clustering. ”, In Proceedings of ACM SIGKDD, pp.126-135, 2006.
[15] L.Baker, A.McCallum, “Distributional clustering of words for text classification. ”, In Proceedings of ACM SIGIR, 1998.
[16] Y.Xue, C.S.Tong, W.S.Chen, W.Zhang, Z.He, “A modified non-negative matrix factorization algorithm for face recognition”, in Proc. Int. Conf. Pattern Recognintion (ICPR), Hong Kong, pp.495-498, 2006.
[17] 亀岡弘和，ルルージョナトン，“Frobenius ノルム規準の非負値行列因子分解における乗法更新式に関する一考察”，日本音響学会講演論文集，pp.709-712，2009.